# An Information Maximization approach of ICA for Gender Classification

[1] Sunita Kumari , [2]Banshidhar Majhi
National Institute of Technology, Rourkela, India
[1] sunita.nitrkl@gmail.com, [2]bmajhi.nitrkl.ac.in

***Abstract*—In this paper, a novel and successful method for gender classification from human faces using dimensionality reduction technique is proposed. Independent Component Analysis (ICA) is one of such techniques. In the current scheme, a thrust is given on the different algorithms and architectures of ICA. An information maximization ICA is discussed with its two architecture and compared with the two architectures of fast ICA. Support Vector Machine (SVM) is used as a classifier for the separation of male and female classes. All experiments are done on FERET database. Results are obtained for the different combinations of train and test database sizes. For larger training set SVM is performing with an accuracy of 98%. The accuracy values are varied for change in size of testing set and the proposed system performs with an average accuracy of 96%. An improvement in performance is achieved using class discriminability which performs with 100% accuracy.**

***Index Terms*—infomax, fast, ICA, ICA-1, ICA-2**

## I. Introduction

Face biometric has gained popularity because it is one of the most communicative part of social interactions. Human-Computer Interaction as well needs face recognition, gender recognition, and estimation of age for passive surveillance. However, during identification of any individual the False Acceptance Rate (FAR) increases as the database size increases [1]. Further the time required increases exponentially as we need 1: 5ØAÜ comparisons for identification. Hence, more thrust is put in any biometric system to reduce the search time without compromising with accuracy. In a face recognition system, partitioning the database based on gender is an important research direction. The major challenge forgender classification is two-fold, (1) to find discriminating features and (2) an optimum hyperplane for separating male and female classes. Hence, most researchers thrust on these two issues while attempting gender classification using face.

There are several methods for dimensionality reduction like Principal Component Analysis (PCA), Factor Analysis (FA), Independent Component Analysis (ICA) and many others [2]. PCA has been first applied by Kirby and Sirovich [3] and it is shown that it is an optimal compression scheme that minimizes the mean square error (MSE) between the original and reconstructed image. Later, authors in [4] have shown that it is the most popular method for face recognition. ICA is a generalization of PCA. This field is contradictory with respect to the performance of its algorithms. For face recognition, Bartlett et al. [5], Yuen and Lai [6], and Liu and Wechsler [7] have claimed that ICA outperforms PCA while Baek et al. [8] claims that PCA performs better than ICA. Moghaddam claims that there is no significant difference between PCA and ICA [9]. It carries same effect in case of gender classification also.

There exist various algorithms of ICA like Jade algorithm, infomax ICA, fastICA [10]. In [11], a technique based on fastICA with Support Vector Machine (SVM) is developed for gender classification with an accuracy of 95.67%, but this approach lacks with the description of its architecture. In this paper, a gender classifier based on SVM is developed which deals with infomax and fast ICA with its two architecture.The rest of the paper is organized as follows: Section II provides a description of the feature extraction techniques used. The proposed system is discussed in Section III. Section IV has been presented with simulation results and comparative analysis. Finally, Section V gives concluding remarks.

## II. Dimensionality reduction using independent component analysis

The present work has been motivated by the keen interest of researchers in face recognition using infomax ICA [5]. ICA gives independent data which can be defined in terms of probability density functions (PDFs). Two variables $y1$ and $y2$ are said to be independent if the joint PDF of $y1$ and $y2$ will be factorizable in terms of their marginal PDF and can be achieved by either *minimizing mutual information or by maximizing non-gaussianity.*

ICA is defined as:

For the given set of input sample $x$, ICA finds a linear transform in such a way that $s = Wx$

Such that the components $s$ are as independent as possible. Here, $w$ is an unknown separating matrix and needs to be determined. There exists several algorithms for determining $w$ like Jade, Information Maximization (infomax) and Fast fixed point (fast) ICA. Proposed scheme uses infomax and fast ICA algorithm for finding separating matrix. These algorithms are discussed in the following subsections.

### *A. Fast Fixed Point ICA*

This algorithm is based on maximizing non-Gaussian property of the estimated sources and is measured with the help of differential entropy $J$ called as negentropy.

$$J(y) = H(y_{Gauss}) - H(y)$$

Where, $H(y_{Gauss})$ is the entropy of the gaussian random




DOI: 01.IJIT.02.02. 99

variable and $H(y)$ is the entropy of the observed random variable. Since the gaussian random variables has the most entropy among all random variables, maximizing $J$ leads to extracting sources as independent as possible.

### B. Information Maximization ICA

Infomax is a gradient based neural network and it maximizes information from input to the output network as proposed by Bell and Sejnosky [12]. The information maximization is achieved by maximizing the joint entropy of transformed vector $z = g(W, x)$ where $g$ is a sigmoidal function. The joint entropy is:

$$H(y) = -E[\ln f(y)]$$

Where $f(y)$ is the multivariate joint density function of $y$.

$$f(y) = \frac{f(x)}{|J_W|}$$

Here, $|J_W|$ denotes the absolute value of Jacobian matrix $J_W$, which is defined as

$$J_W = \det\left[\frac{\partial y_i}{\partial x_i}\right]$$

On combining the above equations, $H(y)$ can be written as:

$$H(y) = E[\ln|J_W|] + H(x)$$

Maximization of $H(y)$ can be achieved by adapting $W$ and can be achieved using only the first term.

### C. Preprocessing of the data

Two preprocessing techniques are required to make the problem of ICA estimation simpler and better conditioned, namely,

1) *Mean/ Centering*: It gives the zero–mean data like for an input vector $x$ subtract the mean $(m = Ex)$ from it.

2) *Sphering/Whitening*: It is a linear transformation method which gives the whiten signal $\tilde{x}$ from an input signal $x$ such that its components are uncorrelated and their variances equal unity. Unit variance can be achieved by making $\cdot\, E(\tilde{x}\tilde{x}^T) = I(IdentityMatrix)\,\cdot$

Whitening is done by $E\{xx^T\} = ADE^T$.

Here, $E$ is the orthogonal matrix of eigenvectors of $E\{xx^T\}$ and D is the diagonal matrix of its eigenvalues. The values of $\tilde{x}$ is determined as $\tilde{x} = ED^{-1/2}E^T x$

### D. Architecture of ICA

The goal is to find a better set of basis images so that it can easily estimate an unknown dataset. There are two ways to achieve this goal:

1) *Architecture–1/ ICA–1*: It produces a set of statistically independent basis images. The data matrix $X$ is constructed in such a way that each row represents an individual image. This architecture deals about independence of images so that images are considered as random variables and pixels as trials as shown in Fig. 1.

Two images $X_i$ and $X_j$ are independent if it is not possible to predict the value of an image $X_j$ from $X_i$ by moving across pixels. A set of coefficients can be determined by the linear combination of the independent basis images from which each individual image can be easily constructed.

2) *Architecture–2/ ICA–2*: In ICA–1, the basis images were statistically independent from each other but the coefficients were not but in Architecture–2, the coefficients are statistically independent. It finds a factorial face code for the set of face images. The data matrix *X* is constructed in such a way that each column corresponds to an individual image. This architecture deals with independence of pixels such that pixels will become random variable and images as trials. Two pixels $i$ and $j$ are said to be independent, if by moving across set of images it is not possible to predict the value of $j$ from pixel $i$ as shown in Fig. 2.

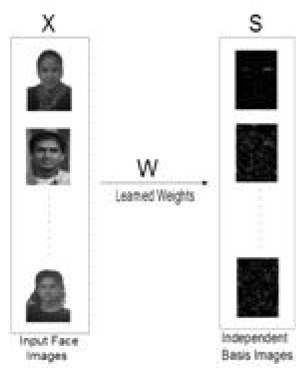


Fig 1. Architecture-1

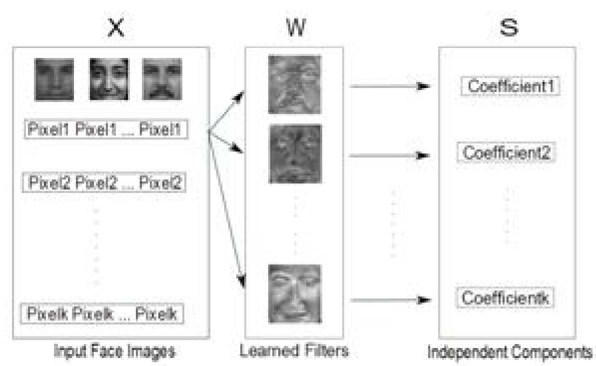


Fig 2. Architecture-2



DOI: 01.IJIT.02.02.99

## III. Proposed gender classification system

Proposed methodology for gender classification is shown in Fig. 3. The performance of gender classification system has been evaluated for two ICA architectures using infomax algorithm and fastICA on the FERET database [13]. Sample images from this database are shown in Fig. 4.

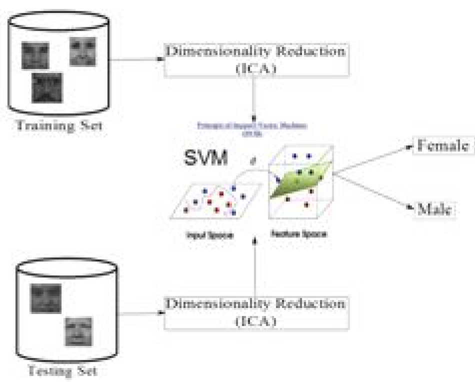


Fig 3. Proposed gender classification system

The dataset consists of images from 500 subjects of which 250 are male and rest is female. To train the SVM, 200 images have been randomly selected from this dataset. From the remaining 300 images, four groups of testing dataset have been made. First test dataset consists of 50 individuals, second has 100 individuals, third consists of 150 and fourth has 200 individual images. Performance of the proposed system is shown on all these test datasets using same training set. The original size of the image was 384×256. All these images have been aligned first and then conventional triangularization approach is applied to get region of interest as shown in Fig. 5.

Features have been extracted using infomax and fast ICA from the cropped images of size 293×241. The size of data matrix is 500×70613, where each row corresponds to one particular image. The number of independent components (ICs) obtained using ICA algorithm corresponds to the dimensionality of the input. As the performance merely depends on the number of separated components so there is a need to control the number of ICs. To achieve this, ICA is performed on linear combination of original images and for this, PCA has used 200 eigenvectors. Thus, ICA has 200 ICs. Basis images obtained from PCA is shown in Fig. 6 and of infomax ICA–1 is shown in Fig. 7. Fig. 8 shows the same as obtained from infomax ICA–2.

Fig 4. Images from FERET database

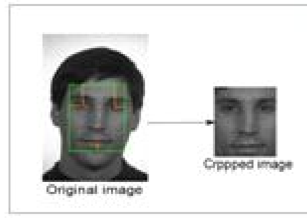


Fig 5. Conventional Triangularization Approach

Fig 6. Basis images from Principal Component Analysis

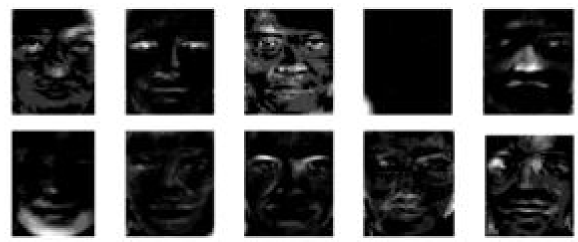

Fig 7. Basis images from ICA Architecture-1: Statistical Independent Images (infomax ICA)

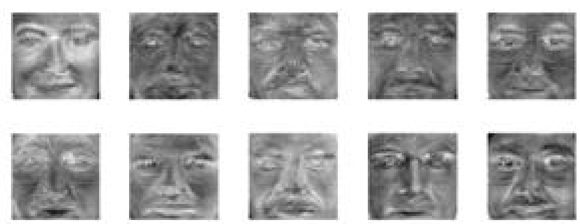

Fig 8. Basis images from ICA Architecture-2: A factorial Face Code (infomax ICA)

### A. Confusion Matrix

The performance of any classifier can be expressed in terms of confusion matrix [14]. Confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems can be evaluated using data present in the matrix and can be defined in the terms of True Positive (TP) and False Positive (FP) rate. The tabular representation of confusion matrix is as shown in Table 1.

Table I. Confusion matrix

| Predicted / Actual | Class-1 | Class-2 |
|---|---|---|
| Class-1 | a | b |
| Class-1 | c | d |


DOI: 01.IJIT.02.02.99

The entries in the confusion matrix are as following:

- a is the number of correct predictions that an instance is negative.
- b is the number of incorrect predictions that an instance is positive.
- c is the number of incorrect predictions that an instance is negative.
- d is the correct number of predictions that an instance is positive.

The performance of confusion matrix can be measured in the following terms:

- *True positive rate:* It is the proportion of positive cases that are correctly classified. $TP = a / c + d$
- *False positive rate:* It is the proportion of negative cases that are incorrectly classified as positive. $FP = b / a + b$
- Accuracy: It is the total number of predictions that are correct.

$$Accuracy = (a + d) / (a + b + c + d)$$

## IV. Experimental results

### A. Results from infomax ICA

The distribution of TP and FP rates and corresponding accuracies for different datasets using infomaxICA for its two architectures is shown in Table 2 and 3. From table, it is clear that the male and female classes are more separable in case of architecture–1 in comparison to architecture–2.

Thus, statistical independent images perform better than statistical independent coefficients. As literature entails that ICA–1 and ICA–2 gives local and global features respectively, thus, the proposed gender classifier is performing well with local features than compared to global for separation of male and female classes.

Table II. Confusion Matrix for Information Maximization ICA Architecture - 1

| Testdata(50)-Traindata(200) | | | Testdata(100)-Traindata(200) | | |
|---|---|---|---|---|---|
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 25 | 0 | Male | 49 | 1 |
| Female | 1 | 24 | Female | 3 | 47 |
| TP = 0.96, FP = 0<br>Accuracy = 98% | | | TP = 0.94, FP = 0.02<br>Accuracy = 96% | | |
| Testdata(150)-Traindata(200) | | | Testdata(200)-Traindata(200) | | |
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 71 | 4 | Male | 94 | 6 |
| Female | 6 | 69 | Female | 9 | 91 |
| TP = 0.92, FP = 0.053<br>Accuracy = 93.33% | | | TP = 0.91, FP = 0.06<br>Accuracy = 92.5% | | |

Table III. Confusion Matrix for Information Maximization ICA Architecture - 2

| Testdata(50)-Traindata(200) | | | Testdata(100)-Traindata(200) | | |
|---|---|---|---|---|---|
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 25 | 0 | Male | 48 | 2 |
| Female | 2 | 23 | Female | 4 | 46 |
| TP = 0.92, FP = 0<br>Accuracy = 96% | | | TP = 0.92, FP = 0.04<br>Accuracy = 94% | | |
| Testdata(150)-Traindata(200) | | | Testdata(200)-Traindata(200) | | |
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 70 | 5 | Male | 91 | 9 |
| Female | 8 | 67 | Female | 13 | 87 |
| TP = 0.89, FP = 0.067<br>Accuracy = 91.33% | | | TP = 0.87, FP = 0.09<br>Accuracy = 89% | | |

### B. Results from fast ICA

In this section, an emphasis is given on fast ICA using architecture–2 and is compared with architecture–1. Table 4 and 5 describes architecture–1 and architecture–2 respectively. FastICA performs better with architecture–2 in comparison to architecture–1. Thus, global features are given more weight in case of fast ICA for gender classification.

Table IV. Confusion Matrix for fast fixed point ICA Architecture-1

| Testdata(50)-Traindata(200) | | | Testdata(100)-Traindata(200) | | |
|---|---|---|---|---|---|
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 25 | 0 | Male | 48 | 2 |
| Female | 2 | 23 | Female | 5 | 45 |
| TP = 0.92, FP = 0<br>Accuracy = 96% | | | TP = 0.90, FP = 0.04<br>Accuracy = 93% | | |
| Testdata(150)-Traindata(200) | | | Testdata(200)-Traindata(200) | | |
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 70 | 5 | Male | 92 | 8 |
| Female | 9 | 66 | Female | 15 | 85 |
| TP = 0.88, FP = 0.067<br>Accuracy = 90.67% | | | TP = 0.85, FP = 0.09<br>Accuracy = 88.5% | | |

Table V. Confusion Matrix for fast fixed point ICA Architecture - 2

| Testdata(50)-Traindata(200) | | | Testdata(100)-Traindata(200) | | |
|---|---|---|---|---|---|
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 25 | 0 | Male | 49 | 1 |
| Female | 1 | 24 | Female | 3 | 47 |
| TP = 0.96, FP = 0<br>Accuracy = 98% | | | TP = 0.94, FP = 0.02<br>Accuracy = 96% | | |
| Testdata(150)-Traindata(200) | | | Testdata(200)-Traindata(200) | | |
| Actual \ Predicted | Male | Female | Actual \ Predicted | Male | Female |
| Male | 71 | 4 | Male | 93 | 7 |
| Female | 7 | 68 | Female | 10 | 90 |
| TP = 0.907, FP = 0.053<br>Accuracy = 92.67% | | | TP = 0.9, FP = 0.07<br>Accuracy = 91.5% | | |


DOI: 01.IJIT.02.02.99

The separation between true positive and false positive rates is described in Fig. 9 for both algorithms and architecture of ICA. The graph shows the relationship between TP and FP against change in test database size. It shows that, the separation between TP and FP is more for infomaxICA–1 than others scheme. From results, it has been inferred that fast ICA– 2 performs better than fast ICA–1.

Based on their TP and FP rates, the corresponding drawn Receiver Operating Characteristic (ROC) graph is shown in Fig. 10. ROC graph defines the tradeoff between the ability of a classifier to correctly identify positive cases and the number of negative cases that are incorrectly classified. The ROC curve shows the plot of false positive rate against true positive rate. A point (0, 1) in the ROC graph shows perfect classification which classifies all positive and negative cases correctly.

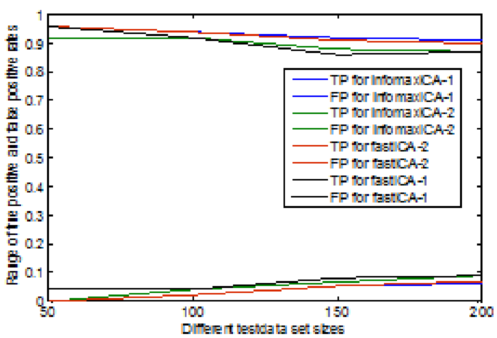


Fig 9. Separation between true positive and false positive rates for different datasets

The point (0, 0) represents a classifier which predicts all cases as negative. FP and TP values are obtained for all testdata sets and ROC is plotted against them. From above graph, it can be said that infomax classifier work better than others.

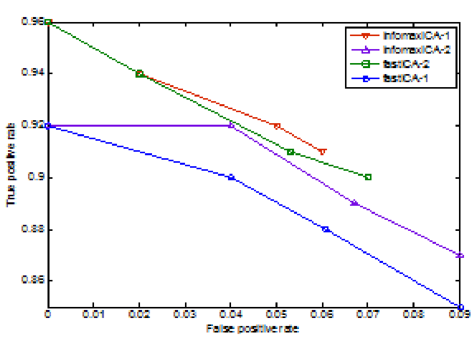


Fig 10. Receiver operating characteristic (ROC) graph

### C. Comparative analysis

Author in [4] has evaluated the performance of fastICA in Architecture–1 with an accuracy of 95.67%. For comparison issues, the proposed system has also evaluated the performance using fastICA architecture–2 and then it is compared with the results obtained from infomax ICA–1, infonax ICA–2, fastICA–1 and PCA based scheme as shown in Fig. 11. By comparative analysis, it has been observed that infomax ICA– 1 performs superior than all other schemes and it is almost similar to fast ICA–2. The performance of infomax ICA–2 and fast ICA–1 is also similar and all these techniques performs better than PCA. Thus it is observed that both the algorithms are giving same performance but it merely depends on the architecture. Infomax ICA performs better in architecture–1 and fast ICA performs better in architecture–2.

### D. Mutual information

Variation of accuracy achieved in case of ICA–1 and ICA–2 depends on the mutual information conserved in their basis images as it is a measure of statistical dependence. Mutual information is calculated as

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

Where $H(X)$ and $H(Y)$ are the marginal entropies and $H(X,Y)$ is the joint entropy of $X$ and $Y$.

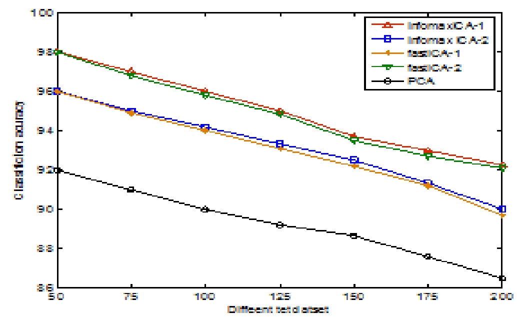


Fig 11. Comparison of proposed scheme with existing Schemes

Two similar images contains high mutual information while mutual information between dissimilar images are low. Such observation is also found in the current approach. A comparison of mutual information between original gray level images and basis images obtained from ICA is shown in Fig. 12.

This graph shows the mutual information present among various images and their basis images. This is obtained using 10 images and the mutual information with respect to other images are calculated. The mutual information present in ICA–1 is less than that of ICA–2 and the original gray level images have greater mutual information than ICA–2. Mutual information is inversely proportional to the independence of the data.Thus ICA–1 have more independence than ICA–2 and original images. For this reason, ICA–1 performs better than ICA–2.

### E. Feature Sub-space Selection

All results are obtained using 200 basis vectors but among all these coefficients, some of them are less significant than others. Thus, an improvement in performance can be achieved by discarding those less significant coefficients while training the SVM. One of the most popular approach for coefficient selection is by maximizing the ratio of between class to within class variance which is also known as class discriminability (cd) and can be obtained as

$$cd = \sigma_{between} / \sigma_{within}$$


DOI: 01.IJIT.02.02. 99

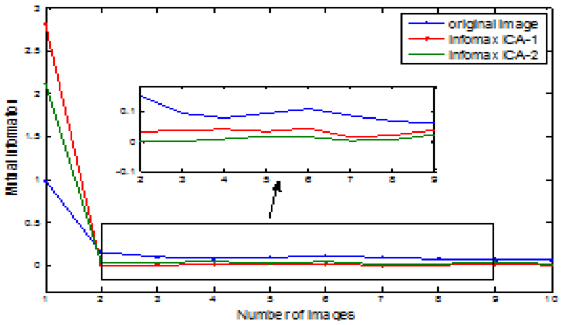


Fig 12. Comparison of mutual information between original and basis images from ICA–1 and ICA–2(this is obtained with 10 basis images where mutual information of first image with itself is calculated first and then it is calculated with other images.)

Where $\sigma_{between}$ is the variance between male and female classes and $\sigma_{within}$ is the sum of variances within each class.

The coefficients with low 5ØPÜ5ØQÜ values are discarded and SVM is trained with the rest. In the current scheme the cd coefficient's value lies between 0.00008635–6.6646. An analysis has been made using these coefficients and is shown in Fig. 13. It is inferred that if the coefficients are selected with the value greater than 0.15 or 0.2, accuracy reaches to 100%. However, the accuracy degrades further if the value is increased as shown in Fig. 13.

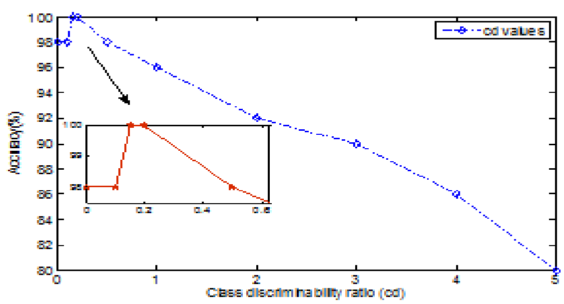


Fig 13. Performance measure with sub-space selection

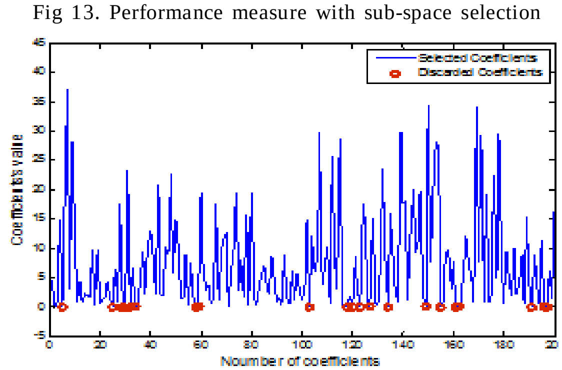


Fig 14. Class discriminabilty between male and female classes (each solid line represents the coefficient's cd value and circle lists the discarded coefficients)

With the implementation of above scheme, the classifier performance has increased significantly. The accuracy obtained is 100% for infomax ICA–1 with 50 testdata and 200 traindata size. Thus using such feature selection schemes, the classifier performance is increased. So, gender classification system is acceptable in reducing the search space for a real-time face recognition system.

## Conclusions

This paper proposes a gender classification scheme based on ICA. Two well-known algorithms for ICA are infomax ICA and fast ICA. SVM is used as a classifier. The performance of the SVM has been studied using different architectures namely, ICA–1 and ICA–2. Comparative analysis has been made with the existing PCA based classification scheme. It has been found that infomax ICA results better in architecture–1 with 100% accuracy and fast ICA performs better in architecture–2 with an accuracy of 100%. Both the algorithms outperform PCA. Thus, gender classification using human faces can be taken as a trusted step while dealing with the large databases. This can be used for filtering databases during identification.

## References

[1] A. K. J. D. Maltoni, D. Maio and S. Prabhakar, Guide to Biometrics, Springer Verlag 2003.

[2] K. K. Sergios Theodoridis, Pattern Recognition, Springer-Verlag 2006.

[3] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A*, 4(3):519–524, Mar 1987.

[4] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[5] M. Bartlett, J. Movellan, and T. Sejnowski. Face recognition by independent component analysis, Neural Networks, IEEE Transactions, Neural Networks, IEEE Transactions, 13(6):1450-1464, Nov-2002

[6] P. C. Yuen and J. H. Lai. Independent component analysis of face images. In *IEEE Workshop on Biologically Motivated Computer Vision*, 2000.

[7] C. Liu and H. Wechsler. Comparative assessment of independent component analysis (ica) for face recognition. In *International Conference on Audio and Video Based Biometric Person Authentication*, pages 22–24, 1999.

[8] K. Baek, B. A. Draper, J. R. Beveridge, and K. She pca vs ica: A comparision on the feret dataset.Algorithm. JOINT CONFERENCE ON INFORMATION SCIENCES. pages 824–827,2002

[9] B. Moghaddam, Principal manifolds and Bayesian subspaces for visual recognition. In computer vision, The proceedings of the 7th IEEE international conference. Pages 1131-1136 vol 2.

[10] A. Hyvrinen and E. Oja. Independent component analysis: algorithm and applications. Neural Networks, 13:411430, 2000

[11] A. Jain and J. Huang. Integrating independent components and support vector machines for gender classification. Volume 3, pages 558 – 561, Aug 2004..

[12] A.J. Bell and T. J. Sejnowski An information maximization approach to blind separation and blind deconvolution. NEURAL COMPUTATION, 7:1129–1159, 1995.

[13] P.J.Phillips H.Moon, S.A.Rizvi and P.Z.Rauss. The FERET evaluation methodology for face recognition.

[14] http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix.html


DOI: 01.IJIT.02.02.99